IMPLEMENTATION ARTICLE

Related to other papers in this special issue	3 (p30); 20 (p199); 7 (p66)
Addressing FAIR principles	F1, F2, F3, F4, A1, R1

The "A" of FAIR – As Open as Possible, as Closed as Necessary

Annalisa Landi[1†], Mark Thompson[2], Viviana Giannuzzi[1], Fedele Bonifazi[1], Ignasi Labastida[3], Luiz Olavo Bonino da Silva Santos[4] & Marco Roos[2]

[1]Fondazione per la Ricerca Farmacologica Gianni Benzi Onlus, 30 – 70010 Valenzano (BA), Italy

[2]Leiden University Medical Center, Leiden, 2333 ZA, The Netherlands

[3]Learning and Research Resources Centre (CRAI), Universitat de Barcelona, Catalunya 08007, Spain

[4]GO FAIR International Support & Coordination Office (GFISCO), Leiden, The Netherlands

Keywords: Accessibility of health data; FAIR metadata; GDPR; Data sharing; ELSI

Citation: A. Landi, M. Thompson, V. Giannuzzi, F. Bonifazi, I. Labastida, L.O. Bonino da Silva Santos & M. Roos. The "A" of FAIR – As open as possible, as closed as necessary. Data Intelligence 2(2020), 47–55. doi: 10.1162/dint_a_00027

ABSTRACT

In order to provide responsible access to health data by reconciling benefits of data sharing with privacy rights and ethical and regulatory requirements, Findable, Accessible, Interoperable and Reusable (FAIR) metadata should be developed. According to the H2020 Program Guidelines on FAIR Data, data should be "as open as possible and as closed as necessary", "open" in order to foster the reusability and to accelerate research, but at the same time they should be "closed" to safeguard the privacy of the subjects. Additional provisions on the protection of natural persons with regard to the processing of personal data have been endorsed by the European General Data Protection Regulation (GDPR), Reg (EU) 2016/679, that came into force in May 2018. This work aims to solve accessibility problems related to the protection of personal data in the digital era and to achieve a responsible access to and responsible use of health data. We strongly suggest associating each data set with FAIR metadata describing both the type of data collected and the

[†] Corresponding author: Annalisa Landi (E-mail: al@benzifoundation.org, ORCID: 0000-0001-9368-6424).

accessibility conditions by considering data protection obligations and ethical and regulatory requirements. Finally, an existing FAIR infrastructure component has been used as an example to explain how FAIR metadata could facilitate data sharing while ensuring protection of individuals.

1. INTRODUCTION

"Recent rapid expansions of health and biological data, combined with the availability of sophisticated computational technologies, offer unprecedented opportunities to benefit public health", as discussed during a workshop held by the European Medicines Agency on November 2016 [1]. Significant scientific insights and findings may derive from the matchmaking of human experts' knowledge with the capability of computers to analyze and share the large amount of data.

Health data are of real value for scientific research and there is an urgent need to reconcile the benefits of data sharing with privacy rights and constraints and ethical and regulatory requirements. Findable, Accessible, Interoperable and Reusable (FAIR) metadata[①] should be able to provide responsible access to health data by considering both the safeguard of the subjects and the need for sharing data that may be used as a basis for decision-making when converted into knowledge and by improving sophisticated computational technologies [2]. The FAIR principles support the idea that data are a resource to be used to formulate and test scientific hypotheses and aim to increase the worth of data by enabling researchers to reuse existing data. In fact, it should be considered that, by increasing the linkage between different types of data (e.g., electronic health records, genetic data, patient-reported outcomes, data derived from clinical trials and studies) and their reuse, a lot of benefits could be produced including an increase of disease knowledge, an earlier diagnosis, a better choice of the treatment – possibly personalized – and a major involvement of patients in the decision process [3].

In the last years, "Interoperability" was considered the bottleneck of the FAIRification process [4], but nowadays it is well acknowledged that another major concern regards "Accessibility", especially if sensitive data are processed. The current growing concerns for data accessibility are mostly related to its connection with "Reusability". In fact, one of the main foci of the FAIR principles is ensuring that research data are reusable in order to become as valuable as possible and to speed up scientific research.

Nevertheless, despite the significantly increased scale of health data processed and available to be reused, the huge spread of the FAIR principles concept and application and the rapid technological development that could lead to fast scientific advances, challenges for the protection of personal data exist and need to be faced. Since the new European General Data Protection Regulation (GDPR) came into force

[①] Metadata are descriptive information about the context, quality and condition, or characteristics of the data.

in May 2018 [5], additional obligations on the protection of natural persons with regard to the processing[2] of personal data[3] have been added to the existing ones. Sometimes, even when data usage licences allow data sharing [6], access to data is not guaranteed for many other reasons. Most of the restrictions may derive from the original consent obtained by the data owner, from data protection policies of the involved institutions, from data sharing/use agreements between data provider and data recipient and often big efforts are required to transform data in a machine-readable format.

Notably, considering the growing concern about the accessibility of health data, there is a number of initiatives aiming to properly address data protection and ethical and regulatory issues regarding data sharing and access to health data and to promote a major involvement of patients in the decisional process. For instance, in the framework of the training activities on the GDPR, the Luxembourgish node of ELIXIR, the European life-sciences Infrastructure for biological Information, is developing the open source software Data Information System (DAISY) for a research data registry. DAISY will allow data owners to create their own Data Information Sheets including the essential data protection metadata, such as data use restrictions, de-identification methods and the legal basis for the processing [7]. Responsible sharing of biomedical data and biospecimens via the "Automatable Discovery and Access Matrix", the ADA-M initiative [8], released by the Global Alliance for Genomics and Health "GA4GH" and the International Rare Diseases Research Consortium "IRDiRC" in 2016, allows data owners to choose the level of visibility or the degree to which different data types are available. Data owners may establish the possibility to be re-contacted in case of new findings and describe through metadata the accessibility conditions (e.g., limitations and permissions) of their "profiles" by expressing preferences regarding data users [9].

A similar initiative, the Beacon network[4] [10], an open sharing platform developed by GA4GH and ELIXIR, allows data owners to choose the access level (open, registered or controlled) to publish their genomics data and define which type of data may be provided to each type of data requestors.

2. FAIR METADATA DESCRIBING ACCESSIBILITY CONDITIONS

The possibility to automate the accessibility conditions as well as compliance with regulatory and ethical requirements and the data protection obligations need to be addressed in order to define the right balance between the safeguard of individuals' privacy and the potential benefits that may be generated by automated machine processing and reuse of data for research.

[2] GDPR definition: "processing" means any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

[3] GDPR definition: "personal data" means any information relating to an identified or identifiable natural person ("data subject"); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person.

[4] https://beacon-network.org/.

In order to solve accessibility challenges and to achieve a responsible access to and use/reuse of data, a scientific multidisciplinary approach needs to be adopted involving different healthcare professionals, researchers, technical and ethical/regulatory and legal experts in the setting up of FAIR metadata.

The H2020 Program Guidelines on FAIR Data (H2020 Program) introduced the concept that "*data should be as open as possible and as closed as necessary*" [11]. In particular, the letter "A" in FAIR stands for "*Accessible under well-defined conditions*". The Accessibility principle[5], included in the FAIR principle, foresees the creation of metadata describing all the conditions for responsible access to and responsible use of data. According to the FAIR principles, the metadata schema should be standardized to simplify the process among data sources and should be able to identify data requestors by using authentication and authorization procedures. For instance, in 2016 ELIXIR launched an Authentication and Authorization Infrastructure (AAI) [10, 12] in order to control and manage access rights of data users, to create different access levels and to meet legal obligations in privacy and data protection legislation. Data users may be identified by using their home organization credentials or community or commercial identities (e.g., ORCID, LinkedIn). Thus, in order to achieve a responsible access to data, infrastructures should use FAIR metadata associated with each data set describing accessibility conditions and AAI infrastructures to authenticate and authorize whoever is requesting access.

Furthermore, FAIR metadata should describe accessibility conditions in a clear and plain language and should be public and accessible even when the data are no longer available.

Additionally, considering that the GDPR aims to strengthen the rights of individuals to be better informed about the processing of their data, which should be lawful and fair, and give them greater control over their own data, it is important to verify if existing metadata take into account the original consent and all existing and applicable data sharing and data use agreements and data protection policies. Moreover, it should be investigated if the metadata have already been implemented with the information to be provided to the data subject when processing personal data according to *Articles 13 and 14* of the GDPR, and if all the data subject's rights have been and will be respected.

Information regarding the type of data processing, the current and future purposes of the processing, and the duration of data storage or the criteria to determine how long data will persist must be provided when processing data. Moreover, additional information to be provided to data subjects are: any transfer of personal data to a third country and the appropriate safeguard measures, any automated decision-making (including profiling) and the responsible figures for data processing (the controller[6], the Data Protection Officer "DPO"[7], the recipients of the personal data). Finally, the data subject may be aware of his/her main

[5] A1. (meta)data are retrievable by their identifier using a standardized communications protocol, A1.1 the protocol is open, free, and universally implementable, A1.2 the protocol allows for an authentication and authorization procedure, where necessary, A2. metadata are accessible, even when the data are no longer available.

[6] GDPR definition: "data controller" means the natural or legal person, public authority, agency or other body which, alone or jointly with others, determines the purposes and means of the processing of Personal Data [...].

[7] "data protection officer" means the person identified by the Organization as being the main contact regarding the protection of personal data.

rights (right to request access to or deletion of personal data, the right to request rectification of personal data, the right to restrict the processing or to object to processing, the right to data portability, the right to lodge a complaint with a supervisory authority and the right to withdraw consent) [5].

Particular attention should be paid when processing health data of patients affected by rare diseases, in particular when non-EU countries are involved in the processing. In particular, health data deriving from patients affected by rare diseases should be considered more sensitive (e.g., identifiable) and additional safeguards measures should be taken.

With reference to non-EU countries, the examples described in the Guidelines 3/2018 on the territorial scope of the GDPR [13] should be consulted. In particular, different situations may take place (e.g., the data are processed in non-EU countries while the data controller/processor[®] is established in EU).

Considering that the use of data in the decision-making process is increasing in the scientific framework, it is important to comply with all the applicable ethical and regulatory requirements and with the data protection obligations. Thus, the goal is to create FAIR operational metadata able to describe the accessibility conditions for each data set considering all the applicable data protection and ethical and regulatory requirements. To achieve this aim, the existing regulatory and ethical documentation (e.g., informed consent forms, data sharing/use agreements, privacy policies, etc.) should be converted in a computable/machine-readable and "machine actionable" format, sufficiently structured and optimized to be processed by a computer, with the aim to speed up the process and to improve the data quality while protecting the data subject's rights. This seems to be the best approach to address all the complex ethical and regulatory requirements and obligations. Therefore, in order to realize a responsible data access model, the actors involved in the access process (e.g., data subject, data controller, data processor, DPO, etc.) should be identified in advance as well as their responsibilities and the applicable accessibility conditions (e.g., limitations and permissions) according to the existing informed consent forms, data sharing/use agreements and data protection policies. Roles and responsibilities must always be well-defined when processing personal data and data subjects must be provided with the name and contact details of the data controller and of the "DPO" (if available) [5].

Technologies need to be implemented with mechanisms and technical protocols detailing accessibility and use constraints (such as limitations, obligations and permissions) to be provided to the data users when requiring access to data.

Notably, there is an urgent need to enrich FAIR infrastructures with sections enabling compliance with GDPR in order to guarantee that data access and use is grounded on a legal basis. In order to achieve this goal and to explain how our considerations should be used in practice, an existing FAIR infrastructure component, the FAIR Data Point (FDP) [14], is used here as an example. For a detailed description of the

[®] GDPR definition: "processor" means a natural or legal person, public authority, agency or other body which processes personal data on behalf of the controller.

FDP design, specification and implementation, we refer to the online documentation [15, 16]. Thus, we review all the four main basic conceptual components of the FDP – namely the Metadata Provider, the Data Accessor, the Security Enforcer and the Metrics Gather, in relation to the aforementioned accessibility considerations.

The Metadata Provider, responsible for providing the data accessibility condition (e.g., limitations, permissions, etc.), should be in charge of creating sufficient FAIR metadata for each data set considering the conditions foreseen by the applicable existing informed consent forms, the data sharing/use agreements and the data privacy policies. GDPR requirements should be implemented as well.

Nowadays, the controlled access model foresees the existence of a data access committee in charge of performing a regulatory and ethical check to allow or deny access to and use of data. As mentioned above, one of the best ways to speed up the process would be the automation of this process by converting the documentation in a machine-readable format. Thus, all the applicable data protection, ethical and regulatory requirements should be considered during the creation of the FAIR metadata outlining the whole framework of applicable accessibility conditions in order to be evaluated in an automatic way, structured, combined and made machine-readable. The effort that is required by data access committees should be reduced as a result, leading to a net overall increase in efficiency and cost-effectiveness of data use in situations where accessibility conditions were previously not well defined, data users belong to different institutions, etc.

Furthermore, the Metadata Provider should also be in charge of modifying accessibility conditions according to the requests of data subjects in order to guarantee the respect of their rights. If the data subject would like to withdraw the consent to process his/her data or restrict the processing or erase/modify the data, the Metadata Provider should evaluate the requests according to GDPR requirements (*Articles 15-22 and 34*) and decide how to proceed. A copy of the data in processing should also be made available to the data subject upon request to allow him/her to verify the lawfulness of the processing [5].

The Security Enforcer, acting as a gatekeeper and protector from the access to the data by requestors that are not allowed to access data, should be in charge of allowing or rejecting access to data after the evaluation of the request. In order to perform a complete assessment of the data access request, the security enforcer should consider both the accessibility conditions associated with each data set and the identity and rights of the requestor.

The specific accessibility conditions should be made available to the Security Enforcer by the Metadata Provider as FAIR metadata associated with each data set. In addition, the Security Enforcer should be tightly coupled to an AAI infrastructure, because an authentication and authorization procedure needs to be followed for each access to any data set or individual data item. For different levels of access, data requestors may be identified through an ORCID, a certified email address, the links to his/her online profiles, with a copy of the identification document, through a (federated) AAI based on a validated institutional account, etc.

The Data Accessor, responsible for providing access to the data included in the data set when access has been granted by the Security Enforcer, should provide a machine-readable interface that allows humans and machines to access data.

The Metrics Gather component, by monitoring all components of the FDP, keeps track of all the data access requests (including the numbers and types of request, the date of the request submission, the name and contact details of the requestor, the decision taken, the data shared/accessed, etc.). In addition, mechanisms to collect and show data users feedback on the process should be developed.

Importantly, we describe a possible implementation by the four basic conceptual components of the FDP, but any other FAIR technology should address these access considerations for personal data. Other FAIR infrastructures may consider the same four basic aspects in order to guarantee responsible data sharing and access to data in compliance with all data protection, ethical and regulatory requirements.

3. CONCLUSION

In this paper, we propose to consider the rights of individuals as integral part of implementing FAIR principles for personal data. We described access considerations to give guidance for users implementing FAIR principles, but we also underlined that these ethical and legal considerations are inevitable for personal data processing. For instance, even if making data findable is the primary goal of a FAIRification process, access considerations necessarily apply when personal data are concerned. In our opinion, this should not lead to avoidance, for instance by preferring the adoption of data anonymization techniques, but to describing access policies in human and machine-readable terms, such that access procedures can be automated for efficient large-scale data visiting.

In conclusion, to achieve a responsible access to data, a good balance between the need for data sharing and the protection of data subjects must be considered. Noticeably, in order to create value from health data it is important to bear in mind the ethical, regulatory and data protection issues related to the access to and use of data.

Thus, in order to solve accessibility problems related to the protection of personal data in the digital era and to achieve responsible access to and responsible use of health data by reconciling benefits of data sharing with privacy rights, ethical and regulatory requirements, FAIR metadata describing accessibility conditions need to be developed.

Accessibility conditions described in the FAIR metadata should consider what is stated in the original consent obtained by the data owner and all the ethical and regulatory documentation applicable to each data set.

GDPR obligations need to be complied with as well by checking if the data subject has received all the required information and by ensuring the respect of data subject's rights on data processing.

Unnecessary restrictions should be avoided and access conditions, including all limitations, obligations and permissions, need to be described in a machine-readable way by converting the documentation in FAIR metadata to be processed at run-time and to be checked against access requirements when asking for access to data. These considerations should be implemented in FAIR infrastructures by identifying roles and responsibilities for data processing and by creating FAIR metadata ensuring the respect of data subjects and a responsible access to health data.

Finally, when FAIR metadata are developed in order to facilitate the free flow of data, while ensuring a high level of protection of individuals' privacy, an efficient, informed and responsible access process to data will be made possible.

AUTHOR CONTRIBUTIONS

A. Landi (al@benzifoundation.org) prepared the first draft of the paper. M. Thompson (m.thompson@lumc.nl) and M. Roos (M.Roos@lumc.nl) contributed to set up of the methodology. V. Giannuzzi (vg@benzifoundation.org) and M. Thompson collaborated in drafting the paper. F. Bonifazi (fb@benzifoundation.org), I. Labastida (ilabastida@ub.edu) and L.O. Bonino da Silva Santos (luiz.bonino@go-fair.org) provided a general revision of the paper. M. Thompson and M. Roos provided a final review of the draft.

ACKNOWLEDGEMENTS

Part of this work is funded by the NWA program (project VWData - 400.17.605), by the Netherlands Organization for Scientific Research (NWO), by the European Joint Program Rare Diseases (grant agreement #825575) and ELIXIR-EXCELERATE (H2020-INFRADEV-1-2015-12).

REFERENCES

[1] European Medicines Agency. Workshop on identifying opportunities for "big data" in medicines development and regulatory science. (2016). Available at: https://www.ema.europa.eu/en/events/workshop-identifying-opportunities-big-data-medicines-development-regulatory-science.

[2] S.O.M. Dyke, A.A. Philippakis, J.R. De Argila, D.N. Paltoo, E.R. Luetkemeier, B.M. Knoppers, … & S.T. Sherry. Consent codes: Upholding standard data use conditions. PLOS Genetics. 12(1)(2016), e1005772. doi:10.1371/journal.pgen.1005772.

[3] C. Ohmann, R. Banzi, S. Canham, S. Battaglia, M. Matei, C. Ariyo, ... & J. Demotes-Mainard. Sharing and reuse of individual participant data from clinical trials: Principles and recommendations. BMJ Open 7(2017), e018647. doi: 10.1136/bmjopen-2017-018647.

[4] A. Jacobsen, R. Kaliyaperumal, L.O. Bonino da Silva Santos, B. Mons, E. Schultes, M. Roos & M. Thompson. A generic workflow for the data FAIRification process. Data Intelligence 2(2020), 56–65. doi: 10.1162/dint_a_00028.

[5] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of The European Parliament and of The Council of 27 April 2016 on the protection of natural persons with regard to the processing

of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)]. Official Journal of the European Union. May, 2016.

[6] I. Labastida & T. Margoni. Licensing FAIR data for reuse. Data Intelligence 2(2020), 199–207. doi: 10.1162/dint_a_00042.

[7] GDPR activities. ELIXIR-LU. Available at: https://elixir-luxembourg.org/gdpr-activities.

[8] GA4GH/ADA-M. Available at: https://github.com/ga4gh/ADA-M.

[9] J.P. Woolley, E. Kirby, J. Leslie, F. Jeanson, M.N. Cabili, G. Rushton, ... & A.J. Brookes. Responsible sharing of biomedical data and biospecimens via the "Automatable Discovery and Access Matrix" (ADA-M). npj Genomic Medicine 3(1)(2018), Article No. 17. doi: 10.1038/s41525-018-0057-4.

[10] M. Linden, M. Procházka, I. Lappalainen, D. Bucik, P. Vyskocil, M. Kuba, … & T. Nyrönen. Common ELIXIR service for researcher authentication and authorization. F1000Research 7(2018), 1199. doi: 10.12688/f1000research.15161.1.

[11] European Commission. Directorate-General for Research & Innovation. H2020 Programme Guidelines on FAIR Data Management in Horizon 2020. Version 3.0. 26 July 2016.

[12] ELIXIR AAI: Authentication and authorization infrastructure. Available at: https://www.elixir-europe.org/services/compute/aai (accessed February 17, 2019).

[13] European Data Protection Board. Guidelines 3/2018 on the territorial scope of the GDPR (Article 3) - Version for public consultation. Adopted on 16 November 2018.

[14] L.O. Bonino da Silva Santos, M.D. Wilkinson, A. Kuzniar, R. Kaliyaperumal, M. Thompson, M. Dumontier & K. Burger. FAIR data points supporting big data interoperability. In: M. Zelm, G. Doumeingts & J.P. Mendonça (eds.) Enterprise Interoperability in the Digitized and Networked Factory of the Future. ISTE Press, 2016, pp. 270–79.

[15] FAIRDataTeam/FAIR Data Point metadata specification. GitHub. (2019). Available at: https://github.com/FAIRDataTeam/FAIRDataPoint-Spec/blob/master/spec.md.

[16] FAIRDataTeam/FAIR Data Point design specification. GitHub. (2019). Available at: https://github.com/FAIRDataTeam/FAIRDataPoint-Spec/blob/master/README.md.